

# Investigating Natural Image Pleasantness Recognition using Deep Features and Eye Tracking for Loosely Controlled Human–computer Interaction

Hamed R. Tavakoli<sup>a,\*</sup>, Jorma Laaksonen<sup>a</sup>, Esa Rahtu<sup>b</sup>

<sup>a</sup>*Department of computer science, Aalto University, Espoo, Finland*

<sup>b</sup>*Center for machine vision research University of Oulu, Finland*

---

## Abstract

This paper revisits recognition of natural image pleasantness by employing deep convolutional neural networks and affordable eye trackers. There exist several approaches to recognize image pleasantness: (1) computer vision, and (2) psychophysical signals. For natural images, computer vision approaches have not been as successful as for abstract paintings and is lagging behind the psychophysical signals like eye movements. Despite better results, the scalability of eye movements is adversely affected by the sensor cost. While the introduction of affordable sensors have helped the scalability issue by making the sensors more accessible, the application of such sensors in a loosely controlled human-computer interaction setup is not yet studied for affective image tagging. On the other hand, deep convolutional neural networks have boosted the performance of vision-based techniques significantly in recent years. To investigate the current status in regard to affective image tagging, we (1) introduce a new eye movement dataset using an affordable eye tracker, (2) study the use of deep neural networks for pleasantness recognition, (3) investigate the gap between deep features and eye movements. To meet these ends, we record eye movements in a less controlled setup, akin to daily human-computer interaction. We assess features from eye movements, visual features, and their combination. Our results show that (1) recognizing natural image pleasantness from eye movement under less restricted setup is difficult and previously used techniques are prone to fail, and (2) visual class categories are strong cues for predicting pleasantness, due to their correlation with emotions, necessitating careful study of this phenomenon. This latter finding is alerting as some deep learning approaches may fit to the class category bias.

**Keywords:** Affective image tagging, natural image pleasantness, eye tracking, deep features.

---

## 1. Introduction

An old desirable objective of artificial intelligence (AI) has been the understanding of emotions by machines. Despite the exceptional recent advancements in computer science, emotion understanding still remains a challenging problem for machines. It is one

of the key distinguishing factors between humans and AI. To reduce the gap between human and machine by incorporating emotions, the field of affective computing has been devoted to the emotion recognition and integration in human–computer interaction (HCI) scenarios. While the recognition of the emotional state of the users is an active area of research, the extent of emotion realization does not need to be limited to the user state. The recognition of the scene pleasantness independent of a user is an example of such a case.

---

\*Corresponding author

Email address: [hamed.r-tavakoli@aalto.fi](mailto:hamed.r-tavakoli@aalto.fi)  
(Hamed R. Tavakoli)

Affective computing utilizes the facial expression and/or the posture of user, physiological signals such as electroencephalography (EEG), magnetoencephalography (MEG), skin conductance response, and eye movement in order to decode the user’s state of mind [1]. On the other hand, the perception of scene pleasantness goes beyond user monitoring and can incorporate machine vision techniques to recognize the scene content in order to judge its emotional message [2, 3]. The first problem is well established and studied to provide user dependent emotion estimation, however, the second problem is viewer independent and is yet a challenge.

The ability to recognize the emotional gist of a scene could facilitate more accurate emotional semantic image retrieval [4, 5]. Further potential applications can include image/video content analysis and annotation, affective assessment and refinement of advertisements, multimedia affective rating, etc. Motivated enough by such applications, in this work, we focus on decoding the pleasantness of images. The image pleasantness or *valence* is the indicator of the amount of positive expression or the negative effect of an image. In order to estimate the pleasantness, we employ computer-vision-based techniques and user-in-the-loop methods as well as the combination of the two procedures.

#### *Contributions:*

Our main contribution is the assessment of image pleasantness recognition in a realistic, less constrained human-computer interaction scenario, in which the most recent computer vision state-of-the-art and eye tracking based techniques are challenged. More precisely, our contribution includes:

- We provide a new data set consisting of eye movements of 20 observers for 382 emotional images. Contrary to the existing data sets, which are often relying on high-end expensive devices to record the eye movements, we utilize an affordable inexpensive device. This helps us to be as close to an everyday HCI scenario as possible.
- It is often argued that the lack of mid- and high-level features has an adverse effect on successful

scene pleasantness decoding using conventional features for machine vision methods. The deep convolutional neural networks (CNNs), however, address this issue. We assess the performance of the deep CNNs for image pleasantness recognition and provide a comparison with the existing vision-based techniques.

- We investigate the replicability of valence decoding from eye movements using affordable eye tracker and a less constrained setup. To this end, we use the 95 images, employed by our former study [6], with new eye movement data.

In the rest of this paper, we first overview the related work. Then we introduce a new dataset including stimuli, setup and procedure for recording eye movements, and analyze the statistics of the collected data. In Section 4, we elaborate the basis of a machine learning approach in which classifiers are trained with various visual and gaze-based features in order to decode the pleasantness category of an image. Section 5 contains several experiments to investigate the performance of features using the adopted machine learning approach. Afterwards, we discuss the results, followed by conclusions.

## **2. Related Work**

### *2.1. Psychology*

Psychology has been the driving force of emotion understanding studies and the inspiration of existing HCI methods. The theories on the emotional message of colors [7] are among influential studies that inspired the early works in image pleasantness recognition [8]. The color theory associates red with positive emotional impact, whereas purple is related with negative emotional feelings. The color theory has evolved through the years and emerged in various psychological models of color emotion [9]. The color theory has been applied to the study of natural and abstract scenes, albeit the study of abstract images mostly benefits from it.

In the span of natural images, the *International Affective Picture System* (IAPS) [10] and its related research studies are among the most well-recognized

ones in the computer science community. The span of such studies covers a wide range, including the correlation of motivation and attention with emotions [11], the affect of scene complexity on emotional arousal [12], and sex differences in emotion perception [13].

The psychology have investigated various properties of images and their emotional affect on observers under different theories such as *gestalt* [14]. There exists studies [15] that investigate the effect of curves on human perception, which are considered appealing. Conversely, chaotic texture or angular and diagonal patterns are considered to evoke negative emotions.

Psychological studies also include the study of different physiological signals, e.g., eye movements. These studies mostly demonstrate the influence of emotions on eye movement patterns. The comparison of observers' eye movements on emotionally different images reveals distortion in the patterns during exposure to highly emotive images [16]. The assessment of the performance of emotive stimuli detection by observers is another indicator of emotion influence on the eye movements, which is highly studied in psychology [17, 18].

## 2.2. Visual Features

The early works, based on visual features, are mostly inspired by the color theory. In a semantic information retrieval task, a series of color-based features are used in [8]. The authors decomposed an image into homogeneous regions and computed the hue, brightness, and saturation as well as the position and the size of each region. Then using a framework of compositional semantics, they performed a bottom-up analysis in a two-level hierarchy to combine the visual features into an expressive feature set, which is hypothesized to be equivalent with the high-level information abstraction in the human brain. Eventually, they employed such features in associating images with semantical terms in order to retrieve images with specific emotional gist.

In [19], an emotional image retrieval system was developed using color semantics. They built a dictionary of color to emotion mappings by clustering the color descriptors using a fuzzy clustering scheme and

implementing a fuzzy system in order to relate colors to emotions. The association of each image segment to an emotion in the dictionary is then decided by the amount of the membership of each pixel, calculated using the fuzzy system. This produces a regional semantic descriptor which is augmented by a descriptor based on the average lightness, average saturation, and the average color contrast of the whole image in order to perform emotional semantic queries.

Later, the color emotion was used for both image emotion classification and retrieval in [20]. In fact, they extended the existing methods by enabling submission of a query image instead of emotion scales. To this end, a kd-tree decomposition over a color-based emotion space was used to map a given image to some emotional category. The same idea was later developed using a bag-of-emotions approach [21]. The bag-of-emotions is a histogram of the number of occurrences of particular emotion patterns.

While the aforementioned techniques rely on some psychological prior assumptions [9] to define a color-based emotion space, later studies started learning the association of color and evoked emotions directly from labeled images. For example, [22] applied a bag-of-visual-colors to learn the emotional gist of abstract paintings from the emotional ratings of users. A similar concept was utilized by [23] using a sparse lasso regression technique. We are seeking a similar approach in employing visual features and extending them to include CNNs.

The visual features are not limited to color and can include texture and shape. Thus, it is possible to utilize, for example, a series of holistic features based on Gabor filters [24] and Wiccest features [25], that encode the edginess and texture characteristics, in order to infer the emotional category of images [2]. In [26], it is tried to capture the scene pleasantness from shapes by using shape-based features in emotion category decoding. They exploited features extracted from line segments, angles, continuous lines, and curves.

The emotion elicitation, however, is not limited to the effect of low-level features discussed above. The higher-level semantic content within an image, such as faces, text, animals, objects, the amount of skin

and so on, and their interaction with themselves and the environment, is by far recognized as a crucial missing component [27]. Particularly, in the case of natural images, content-related features matter. Motivated by such observations, [3] studied the role of content-related features. They revealed that various content-related features can contribute to the emotional gist of the scene. Eventually, they argued that semantic content analysis is crucial in the recognition of the affective categories of images.

Recently, deep Convolutional Neural Networks (CNNs) have influenced the computer vision research substantially. There exist numerous successful applications, such as image and scene classification [28, 29], object detection [30], and object segmentation [31], where the state-of-the-art relies on deep CNNs. During the preparation of the final manuscript, we learned that [32] and [33] used CNN features for pleasantness recognition, under the umbrella of sentiment analysis, to infer if an image is pleasant or unpleasant. In other words, they are solving a two-class, positive and negative, classification problem. While [32, 33] apply an end-to-end approach, that is, they learn the CNN features and task together, we utilize a filter bank approach as in [34] because of the relatively small number of images in our data set. There are also other differences in our problem setup: 1) we use a three class classification, *unpleasant*, *neutral*, and *pleasant* paradigm, which in our opinion is more natural, and 2) in our data set, the ground truth emotional labels are based on the diverse scores of *International Affective Picture System* (IAPS) corpus [10] ratings, while the data used by [32, 33] uses a majority vote scheme of at most five people to decide the emotional class categories.

### 2.3. Eye Movements

The automatic detection of the semantic content is often difficult, and sometimes replaced by other means such as observers’ eye movements. For example, [35] applied eye movements as a proxy to identify scene content such as interacting elements. Similarly, a method for localization of affective objects using eye movements was developed in [36].

The eye movements, however, are not only a cue for semantic content decoding. There exist studies

which address the user independent recognition of the pleasantness of images and videos using various physiological signals including eye movements. In these approaches, the eye movement is a signal, gathered unobtrusively from several observers and aggregated in order to tag an image implicitly and independent of a specific user. For example, [37] utilized EEG and eye tracking for affective video tagging in which they relied on features extracted from pupil diameter, gaze distance, and eye blinking obtained from gaze recordings.

In [6], eye movements were applied in order to recognize the emotional message of an image in terms of its pleasantness. To this end, several features were utilized including fixation duration, fixation location, saccade slope, saccade length, saccade orientation, and saccade velocity in order to build a classifier. Eventually, it has been demonstrated that eye movements outperform the bag-of-visual-colors [22] and visual SIFT [38] features in the case of abstract images.

Later, the influence of the features extracted from eye movements and their contribution for image pleasantness recognition were studied in [39]. Utilizing various feature selection techniques, it was demonstrated that the most influential features for such a task are the fixation duration and fixation density map. Furthermore, the analysis of feature representation schemes confirmed that the histogram-based feature representation have the edge over traditional average-value representations.

In regard to the eye movement features, we are extending the works of [6] by incorporating more images with eye movements. There exists a major difference that is on the sensor side and data. Instead of a high-end accurate eye tracker that is not available to everyone, we are utilizing an affordable sensor available to any user. We, further, adopt a less restrictive experiment setup to mimic user interaction in an everyday computer use setting where there is no expert guiding the calibration and users perform the calibration themselves upon training.

## 3. Data Set

The data set consists of affective images and corresponding eye movements of several observers for

each image. The eye movements are recorded in a free-viewing task. We first explain the stimuli and its characteristic. Then we elaborate the experiment setup, procedure, and task. Finally, we analyze the recorded eye movements using conventional statistical analysis approaches.

### 3.1. Stimuli

We have a total of 382 affective images, all selected from the *International Affective Picture System* (IAPS) corpus [10]<sup>1</sup>. The images have the resolution of  $1024 \times 768$ . The image set includes the same 95 images which were previously used in the studies of emotional valence recognition [6]. There is gender-specific emotional rating agreement and only one class of visual content, labeled by human, for the 95 images. The set of 382 images, however, includes a wide span of emotional and visual contents including reptilians, wild predators, domestic animals and pets, people and activities, portraits, erotic and nude, objects (e.g., cup, stool, etc), foods, events (e.g., flood, explosion, protest, etc).

The 382 images are accompanied with the emotional valence ratings in the form of mean self-assessment manikins (SAM) [40] score per image in which the score of 1 indicates the most unpleasant case and 9 is the most pleasant image. The mean SAM score of each images is provided in terms of the genders and all users. Assigning images with mean valence value in the range of 4 – 6 as neutral, the image set consists of 134 pleasant ( $\mu = 6.96, \sigma = 0.61$ ), 84 unpleasant ( $\mu = 2.94, \sigma = 0.64$ ) and 164 neutral images ( $\mu = 5.10, \sigma = 0.55$ ). Based on the gender-specific emotional ratings of the IAPS, there is a strong agreement between genders on the emotional content of only 296 of the images; from which 102 are pleasant ( $\mu = 7.16, \sigma = 0.53$ ), 68 are unpleasant ( $\mu = 2.78, \sigma = 0.56$ ) and 126 are neutral ( $\mu = 5.04, \sigma = 0.47$ ).

<sup>1</sup>The images can not be shown due to copyright restrictions imposed by the image corpus. Please visit <http://csea.php.ufl.edu/media.html> to obtain the images from the center for the study of emotion and attention, University of Florida.

### 3.2. Observers

The participants are 20 volunteers, 12 male and 8 female, with mean observer age of 29.9 (std=7.36, min=20, med=27, max=53). They are graduate and postgraduate students majoring in computer science. The participants have normal or corrected to normal vision and never reported having eye-sight problems, nor any psychological disorders. They have not previously seen the stimuli. Among the images 15 are always displayed twice, i.e., the participants watch 397 images, though the second run of an image is not used in our experiments. Each participant views all the images in one session. The whole session including the instructions does not exceed 1 hour.

### 3.3. Eye Tracking Procedure

We are using a Tobii EyeX eye tracker, which is one of the most affordable devices (less than 100 Euro in 2014), to record the eye movements. The information regarding the configuration and specification of the device is summarized in Table 1. We use the provided API in order to determine the gaze location and fixation events using the default recommended parameters. The images are displayed on a 19 inch LCD at the resolution of  $1600 \times 1200$ . The images are screened at the center of a black background in their original resolution ( $1024 \times 768$ ). Each image is presented for 5 seconds followed by a gray mask for 2 seconds. To be less restrictive, the observers are distanced at about 65–70 cm from the monitor at their convenient sitting posture and no chin rest is used. They are instructed to keep their heads within the tracking plane, which provides the freedom of small head movements, and to watch the images during the viewing time. The calibration procedure is 9-point.

Since we would like to be as close to an uncontrolled environment and mimic unsupervised interactions as in daily life, each observer is initially given instructions on using the eye tracker and helped with the calibration procedure. Once the users know how to use the device, they are left to repeat the calibration procedure once again themselves, as expected in a real-world application scenario, and run the viewer application. The viewer application simply presents the images as described above and records the users' fixation and gaze information.

Table 1: Tobii EyeX Tracking Device Specification and Configuration.

Specification & Configuration	value
Sampling rate	> 60Hz
Latency	15 ms +/- 5 ms
Operating distance	45-100 cm
Headbox size	40 x 30 cm at 65 cm
Firmware	2.0.2-33638
Core version	2.0.9
Driver version	2.0.9
Service version	1.9.4.6493
Engine version	1.9.4.6493
Configuration version	3.2.9.521
Interaction version	2.1.1.3125
C++ SDK version	1.7

### 3.4. Statistics

The data set includes approximately 7 hours and 44 minutes of gaze data consisting of 120219 fixations. 51574 of fixations are on neutral images, 42173 on pleasant, and 26472 on unpleasant images. To gain further insight about the data, we looked into some influential characteristics, which are repeatedly reported of having a role in identifying the affective state of observers and image pleasantness tagging, namely “fixation duration”, “saccade slope”, and “saccade length” [41, 37, 42, 43, 39]. For this purpose, we pulled the data of all the observers together and computed the average feature per image in each emotion category. Figure 1 presents the box plots of the fixation duration, saccade length, and saccade slope for the unpleasant, neutral, and pleasant images along with the number of fixations in each emotion category. Using this data, we did an ANOVA test, which indicates that there is no significant difference ( $p > 0.05$ ) between the three emotional classes in terms of average fixation duration for  $[F(2, 379) = 0.33, p = 0.72]$  and average saccade slope for  $[F(2, 378) = 0.13, p = 0.88]$ . The average saccade length is, however, significantly different ( $p < 0.05$ ) for emotion categories for  $[F(2, 378) = 4.42, p = 0.012]$ . The Tukey-Kramer’s post-hoc test reveals that the saccade length significantly differs for unpleasant images compared with neutral ones, while there is no significant difference between the neutral and pleasant images as well as for the pair of pleasant and unpleasant images.

We also repeated the above analysis after removing the images with strong disagreement between the

genders. The results are similar, i.e., for average fixation duration there is still no significant difference ( $p > 0.05$ ) between the three emotional categories for  $[F(2, 293) = 0.49, p = 0.61]$  and also for saccade slope  $p > 0.05$  for  $[F(2, 292) = 0.56, p = 0.57]$ . The mean average saccade length is significantly different ( $p < 0.05$ ) for emotion categories for  $[F(2, 293) = 4.17, p = 0.02]$ , where the Tukey-Kramer’s post-hoc test again shows the saccade length significantly differs between unpleasant and neutral images with no difference between the pair of neutral and pleasant images as well as the pair of unpleasant images and pleasant ones.

To examine the possible differences between genders [13, 44] in the current data, we pulled the data of each gender category and compared them against one another using ANOVA. The results indicate that for the current data, there is no significant difference ( $p > 0.05$ ) between the genders in terms of average fixation duration for  $[F(588, 1) = 2.27, p = 0.13]$  and saccade slope  $[F(587, 1) = 0.15, p = 0.69]$ . There is, however, a significant difference ( $p < 0.05$ ) between the genders in terms of mean saccade length  $[F(587, 1) = 57.72, p \approx 0.0001]$ , i.e., the saccade length statistic is different between male and female participants per image.

An interesting statistic to check is the variability among subjects in looking at the same images. To measure such a variability, we compute the inter-observer visual congruency [45] for the data. For this purpose, we use fixation locations and adopt a leave-one-out policy as in [45]. We leave out one participant and build a fixation density map from the rest of the fixations. Then, we assess how well the fixations of the left-out participant consists with the rest of the observers using the area under the curve (AUC) metric. For an image, the average of the AUC score of the participants is recognized as the inter-observer visual congruency (IOVC) score of that image. We compute the mean IOVC score for all the images and each image category. The IOVC for all the images is 95.34% +/- 1.9. Looking finer into the images in each emotion class category, the IOVC values are 95.45% +/- 1.5 for the unpleasant, 95.43% +/- 1.7 for the neutral, and 95.16% +/- 1.8 for the pleasant images. Apparently, there is no significant difference ( $p > 0.05$ )

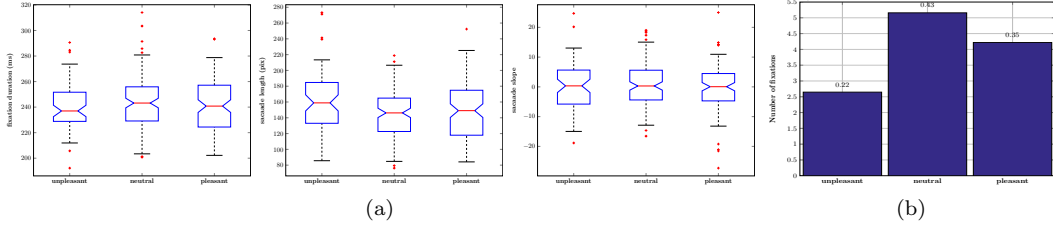


Figure 1: Eye movement statistics: a) box plots for fixation duration, saccade length, and slope, for different emotional categories, b) Number of fixations per emotional category with their ratio to the total number of fixations.

between emotional class categories in terms of mean IOVC of images for  $[F(2, 379) = 1.15, p = 0.32]$ .

We also measure the amount of center bias [46] in the database and in each emotional class category. We gathered all the fixations together to obtain a fixation map. Afterwards, a two-dimensional normal distribution  $\mathcal{N}(\mu, \Sigma)$  is fitted, where  $\mu = [\mu_1, \mu_2]$  is the average location and  $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$  is the covariance of the locations. Figure 2 depicts the center bias for the data set in terms of mean eye positions, the horizontal and vertical projections of the probability distribution of fixations within the normalized image size of  $[-1, 1]$ , where  $(0, 0)$  is the image center. The parameters of the estimated distribution are  $\mu = [0.006, -0.046]$  and  $\sigma = [0.23, 0.26]$ . Therefore, the fixations are a bit deviated upwards<sup>2</sup>. It is not surprising that similar to the previous research in visual perception, the current data is also affected by the center bias phenomenon.

#### 4. Method

Nowadays, the use of machine learning approaches in analyzing physiological signals have become a standard norm and most recent studies of eye movements include some machine learning technique to prove a hypothesis, e.g., [47, 48, 49, 43]. This approach has been also favored in the HCI community for a long time and provides the flexibility of integrating various signals and features into a system. In this section,

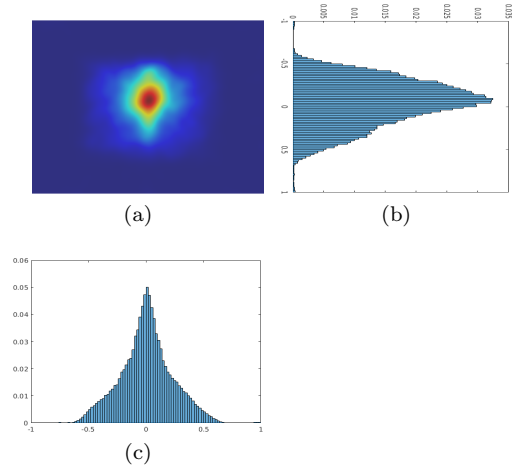


Figure 2: Center bias. a) The mean eye position map (MEP), obtained from smoothing all the fixations of all the participants on all the images. b) The vertical distribution of the fixations. c) The horizontal distribution of the fixations.

we first explain the feature extraction, which includes both features extracted from eye movements and images, followed by the classification scheme.

##### 4.1. Eye Movement Features

It has already been demonstrated that various eye movement features contribute to the recognition of the emotional message of a scene. We employ a selection of features based on aforementioned fixation and saccade characteristics, including fixation locations, fixation duration, saccade length, and saccade slope.

<sup>2</sup>The system coordinates used are compatible with left-handed coordinates as in Computer Graphics.

The fixation location is encoded in terms of fixation density maps. The fixation density map feature is computed by fitting a mixture of Gaussian kernels at the fixation locations. Then the map is down-sampled to the size of  $20 \times 15$  and vectorized to build a feature vector. To avoid scene encoding, which the fixation density map is susceptible to be affected by, akin to [39], we also study the entropy of the fixation density map. The entropy of the fixation density map indicates the amount of information a map carries independent of the exact locations of the fixations.

For the rest of the gaze properties, we use the mean value and standard deviation as the feature representation similar to [47], as well as the histogram representation of [39]. The histogram of fixation duration is computed by quantizing the possible range between the minimum and maximum duration into 60 bins. The number of bins is chosen in relation to time quantiles. Then the number of the occurrence of each bin is counted. A similar approach is adopted to make a histogram of saccade slope and saccade length. The saccade slope histogram consists of 30 bins with quantiles of  $6^\circ$  for the values in the range of  $0^\circ$  to  $180^\circ$ . The histogram of saccade length consists of 50 bins.

As an alternative to the above feature representations, which explicitly exploit the properties of eye movements, we also employ an automated feature learning technique based on Fisher Kernel Learning (FKL) [50]. The FKL algorithm treats the eye movements as a time-series and tries to automatically learn a hidden Markov model representation that maximizes the Fisher criterion. This approach is successfully used for observer task detection from eye movements by [49]. Similar to [49], we employ 10 hidden states and center the data with respect to the training set.

Another fixation-based feature is inter-observer visual congruency. Its emphasis is on the differences between the observers on watching an image. The inter-observer visual congruency have been applied for image ranking [51] and image memorability prediction [52]. We, here, adopt the concept to build a feature vector consisting of the observer visual congruency scores for images. For each image, we compute the consistency of the observers' eye movements

in terms of AUC and report the mean and standard deviation as features to build a 2-dimensional feature vector. It is worth-noting that the proposed feature vector can be seen on as a compressed version of the fixation density map vector which signifies the degree of observer agreement on an image.

#### 4.2. Visual Features

The role of visual elements of a scene can not be neglected. It is often speculated that the computational visual features often fail because they can not fully model the high-level semantical information encoded in an image [27]. Nonetheless, recent advancements in the area of deep neural networks has introduced some hope and expectations for success in the task of image pleasantness recognition. Unlike the engineered features such as SIFT [53], PHOW [54], and SURF [55], a deep network encodes low-, mid-, and some degree of high-level contextual information. In our system, we use visual features including PHOW features and deep features to assess the performance of the deep features in comparison to other visual features and eye movements.

##### *PHOW Features*

Motivated by the characteristic of the SIFT in capturing contours and edges,[22] and [6] applied SIFT descriptors for analyzing the emotional gist of an image. They applied densely extracted RGB-SIFT and RGI-SIFT [56] in order to obtain an image representation.

To incorporate multiple scales, we adopt the Pyramid Histogram of visual Words (PHOW) descriptor [54]. It is a variant of the dense SIFT descriptor which extracts SIFT descriptors at several scales and can be easily adopted to support color features. In particular, we apply the PHOW-color to incorporate color information descriptors. Finally, we encode each image using a Bag-of-Visual-Words (BOW) consisting of a visual vocabulary of 4096 words obtained using a k-means clustering scheme. The vocabulary size is selected not to be too small, nor too large to prevent bad representativeness and overfitting, respectively. The number of 4096 words is the typical vocabulary size for successful and efficient image



recognition tasks. It is worth noting that the PHOW-color variant almost corresponds to the same features used in [22, 6], and [23] and can be considered a replicate of their visual features.

#### Deep Features

Despite the training of CNNs often requires gigantic amount of data, it is possible to apply the features extracted using pre-trained neural networks for a small amount of test data due to the generalization properties of the deep feature extractors [34]. To perform such a domain transfer, the activations on the output of a deep, fully-connected layer is often combined with a target domain specific classifier. The number of images in our data set is relatively small compared to the number of images needed for training a deep CNN or even fine-tuning a pre-trained network for a specific task. Thus, we adopt a bank of filters approach to CNNs [34]. For this purpose, we employ the very deep convolution networks [57], recognized as the VGG architecture, which is shown to be useful in filter bank approaches [34].

The input to the architecture is a fixed-size, mean subtracted image. The image goes through the pre-trained network which consists of convolutional layers with  $3 \times 3$  receptive fields and max-pooling with the stride of 2 pixels between layers. The convolutional layers are followed by two fully-connected layers each having 4096 outputs. We use the output of the last layer as the features in our framework<sup>3</sup>, which is equivalent to a compact feature representation. The performance of the architectures with 16 (VGG16) and 19 (VGG19) layers are studied.

#### 4.3. Classification Scheme

The classification is carried out using support vector machines (SVM) [58]. Following [39], which compared the linear, polynomial, and radial basis SVMs and concluded linear SVM is a proper choice for

<sup>3</sup>Note: the architecture has 3 fully-connected layers, where the last layer performs classification and the other layers can be considered as transformations from the convolution pipeline to a compact feature vector representation. We have reported only the layers utilized.

emotion recognition, we employ linear SVM as the classifier. The classification scheme is one-versus-rest where there exists one classifier for each image pleasantness category. To evaluate the performance, we perform a repeated 10 fold cross-validation, with three sets of *train*, *validation*, and *test* of the ratios of 0.9, 0.05, 0.05, respectively. To deal with the imbalanced data, in each repetition and fold, we guarantee an equal number of samples from each class category following a resampling with partitioning strategy [59] as in [60]. Consequently, the chance level for a three class classification is thus 33.3%. The results are summarized as confusion matrices and mean accuracies (mA). The 95% confidence interval (CI) is also reported. We further perform a McNemar test [61] to evaluate how different the performance of classifiers is from chance.

## 5. Experiments

To conduct the experiments, we consider three scenarios for the stimuli/images. These scenarios are 1) 95 images which have high emotional agreement between genders and narrowing down the image visual class categories to one class category as in [6], that is the visual class of “People & Daily Activity”, annotated by human, 2) 296 images, which are emotionally agreed between the genders, but include any visual class category, 3) the whole 382 image set, where some of the images are not emotionally agreed between genders and a diverse set of visual class categories also exist. Depending on the goal of the experiment, we will use one or all the three scenarios.

### 5.1. Eye Movements

We first perform the experiments with the subset of 95 images used in [6] and learn a classifier to categorize them into three emotional classes of pleasant, neutral, and unpleasant. Figure 2 summarizes the results, showing that none of the features based on eye movements perform significantly above chance level (33%). Then, we extend the image set to the 296 images which are agreed between the genders and learn a classifier to categorize the images into the three possible classes. For this purpose, we use each of the

proposed features individually. The result is summarized in Figure 3. As depicted, it is only the fixation density map that performs significantly better than chance at 37.8% [ $p = 0.01$ , 95% CI=35.65–39.45], with a well-balanced confusion matrix where all the emotional class categories are distinguishable above chance.

We repeated the experiments using the total of 382 images, which includes the images where genders do not agree. Figure 4 summarizes the results. We observe a change in the performance of features, that is, the fixation density map performance can not be distinguished from chance, while the histogram of saccade length starts showing some promising results with accuracy of 37.3% [ $p < 0.001$ , 95% CI=36.27–38.56]. It is worth mentioning that despite the performance of fixation density map is not distinguishable from chance on average, it provides a fair score for all the three classes.

*The combination of features* is also studied for all the three scenarios, where we perform a late fusion. We assessed the late fusion of all the eye movement features along with the fusion of fixation density map and the histogram of saccade slope, which seemed promising for the whole data set in individual feature experiments. The results are summarized in Figure 5, depicting that none of the assessed combination settings does perform significantly above chance.

In this experiment, while the results of the two settings achieving significantly better than chance performance consist with the findings that support the role of fixation patterns and saccade slope (angular behaviour) as reported in [43, 18], we can not achieve a considerable classification score similar to those previously reported by [37] and our former research [6, 39]. Apart from nuance feature and implementation differences, e.g. [37, 39] employ some feature selection techniques, we believe our less restrictive experiment setup also plays a role and hence the current data is more difficult than the data used in earlier research. We will discuss this issue later.

## 5.2. A Finer Look into the Fixation Locations

We encoded fixation locations in terms of fixation density maps. Among the features, by investigating the confusion matrices, the fixation density map

seems performing fairly better than other features in two of the settings. It, particularly, performs significantly better than chance on the subset of 296 images. To summarize, the fixation density map provides the best performance where there exists diverse visual class categories and agreement on emotional gist of an image between genders. This gives rise to the following question, “*Why fixation density map does perform better on this specific setting?*”

The fixation density map can act as a holistic scene representation that provides a reasonable compressed scene descriptor. This is basically the concept behind gist-based [62] operators that exploit saliency for scene recognition like [63]. In other words, it seems the fixation density maps are helping to learn the visual class categories which can potentially be correlated with the emotional gist of the scenes as some visual class categories are biased towards specific emotions, e.g. a photo of food is often identified as neutral or pleasant rather than unpleasant. This explains the reason behind the poor performance of the fixation density maps in the subset of 95 images, where there is only one visual class category available according to human-provided labels [6].

In the case of the 382 images, similarly, the performance is not significantly different from chance. We speculate this is associated with the effect of gender-bias, that is, some of the visual class categories are found emotionally different between the genders [13, 64], in 86 of these images, introduced by incorporating all the images. Such a bias will cause a visual class category to contribute to several emotional class categories and makes the pleasantness recognition from fixation density maps difficult.

We, thus, assess the performance of fixation density maps in terms of entropy to suppress the location information, i.e. we neutralize the role of visual class categories that may have been contributing to the recognition in the 296 images. Figure 6 reports the results of the entropy of fixation density map, which is not significantly better than chance. In other words, discarding the spatial information, which holds a level of the visual category information of scenes, the detection of pleasantness becomes impossible even for the 296 images. Later, we will investigate the role of visual class categories with the help of visual features.

		Prediction			Prediction			Prediction						
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant				
Actual	unpleasant	32.1	33.6	34.3	35.7	34.0	30.3	41.7	32.7	25.6				
	neutral	30.1	35.0	35.0	33.2	31.6	35.2	28.7	34.0	37.3				
	pleasant	36.6	31.9	31.5	30.1	34.4	35.6	32.5	33.0	34.5				
	mA=32.9% (p = 0.86, 95% CI=29.55–36.11)			mA=34.3% (p = 0.69, 95% CI= 30.85–37.80%)			mA=37.16% (p = 0.31, 95% CI= 33.29–39.03%)							
(a)		mean+std fixation duration			(b)		mean+std saccade length			(c)		mean+std saccade slope		
		Prediction			Prediction			Prediction						
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant				
Actual	unpleasant	26.5	38.5	35.0	35.3	27.4	37.2	33.3	33.3	23.3				
	neutral	37.0	27.6	35.4	33.5	34.8	31.7	33.3	33.3	33.3				
	pleasant	36.5	33.7	29.8	31.2	38.0	30.8	33.3	33.3	33.3				
	mA=28.0% (p = 0.05, 95% CI= 24.93–31.06%)			mA=33.5% (p = 0.94, 95% CI= 29.13–37.86%)			mA=33.3% (p = 1.00, 95% CI=33.33–33.33)							
(d)		histogram of fixation duration			(e)		histogram of saccade length			(f)		histogram of saccade slope		
		Prediction			Prediction			Prediction						
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant				
Actual	unpleasant	33.0	33.5	33.5	31.4	34.5	34.0	43.3	31.4	25.3				
	neutral	35.2	29.6	35.2	36.8	30.9	32.3	28.1	33.2	38.7				
	pleasant	31.8	36.9	31.3	31.2	34.9	33.9	29.0	35.3	35.7				
	mA=31.3% (p = 0.44, 95% CI=29.74–34.92)			mA=32.1% (p = 0.62, 95% CI= 28.49–35.50%)			mA=37.4% (p = 0.13, 95% CI= 29.78–44.87%)							
(g)		fixation density map			(h)		IOVC			(i)		FKL		

Table 2: The performance of different eye movement features on the 95 images, where both genders highly agree on emotional gist of images and there is only one visual class category as described in [6]. Chance level is 33%.

		Prediction			Prediction			Prediction			
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
Actual	unpleasant	34.1	33.4	32.5	33.7	33.5	32.8	33.2	33.2	33.6	
	neutral	33.5	33.1	33.5	33.4	32.6	34.0	33.8	33.0	33.3	
	pleasant	32.2	33.5	34.3	33.0	33.7	33.3	33.0	33.8	33.2	
	mA=33.8% (p = 0.70, 95% CI=32.17–35.60)			mA=33.2% (p = 0.94, 95% CI= 32.23–34.21%)			mA=33.1% (p = 0.89, 95% CI= 31.65–34.56%)				
mean+std fixation duration				mean+std saccade length				mean+std saccade slope			
		Prediction			Prediction			Prediction			
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
Actual	unpleasant	31.8	35.1	33.1	36.5	33.4	30.1	33.7	37.0	29.3	
	neutral	32.9	34.1	32.9	33.3	30.4	36.3	33.2	32.7	34.2	
	pleasant	35.0	31.0	33.9	30.1	36.0	33.9	33.3	31.9	34.8	
	mA=33.4% (p = 0.97, 95% CI= 30.48–36.29%)			mA=33.7% (p = 0.82, 95% CI= 30.48–36.29%)			mA=33.7% (p = 0.87, 95% CI=32.12–35.10)				
histogram of fixation duration				histogram of saccade length				histogram of saccade slope			
		Prediction			Prediction			Prediction			
		unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
Actual	unpleasant	42.0	28.5	29.5	33.0	34.8	32.2	35.5	34.6	29.9	
	neutral	28.8	36.4	34.8	31.6	34.1	34.3	32.3	33.7	34.1	
	pleasant	30.7	34.2	35.0	35.7	31.2	33.1	33.4	32.2	34.4	
	mA=37.8% (p = 0.01, 95% CI=35.65–39.45)			mA=33.5% (p = 0.93, 95% CI= 32.01–34.90%)			mA=34.5% (p = 0.70, 95% CI= 32.17–35.61%)				
fixation density map				IOVC				FKL			

Figure 3: The performance of different eye movement features on the 296 images, where both genders highly agree on emotional gists of images, for recognizing the pleasantness of the images. Chance level is 33%.

Actual	Prediction			Prediction			Prediction		
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant
	34.7	32.5	32.9	32.3	35.3	32.4	34.3	32.8	32.9
	32.8	33.3	34.0	34.0	32.5	33.5	34.1	33.0	32.9
	32.4	34.3	33.4	33.5	32.4	34.1	31.6	34.2	34.2
mA=33.8%			mA=33.0%			mA=33.8%			
(p = 0.70, 95% CI=32.68–34.98)			(p = 0.76, 95% CI= 31.19–34.64%)			(p = 0.71, 95% CI= 32.70–34.96%)			
mean+std fixation duration			mean+std saccade length			mean+std saccade slope			
Actual	Prediction			Prediction			Prediction		
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant
	32.4	32.7	34.9	28.0	34.9	37.1	37.3	34.6	28.1
	32.8	34.0	33.2	35.5	35.0	29.5	25.3	31.6	43.2
	34.7	33.4	32.0	35.8	30.3	34.0	26.5	30.4	43.2
mA=32.8%			mA=32.3%			mA=37.3%			
(p = 0.66, 95% CI= 30.31–35.18%)			(p = 0.58, 95% CI= 30.93–34.15%)			(p < 0.001, 95% CI=36.27–38.56)			
histogram of fixation duration			histogram of saccade length			histogram of saccade slope			
Actual	Prediction			Prediction			Prediction		
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant
	35.2	34.6	30.1	33.9	33.6	32.5	33.9	29.7	36.3
	31.7	33.7	34.6	33.5	33.5	33.1	32.1	35.0	32.9
	32.9	31.4	35.7	32.6	32.9	34.4	34.2	33.8	32.0
mA=34.9%			mA=33.9%			mA=33.7%			
(p = 0.24, 95% CI=32.69–37.05)			(p = 0.34, 95% CI= 32.47–35.41%)			(p = 0.86, 95% CI= 30.67–36.48%)			
fixation density map			IOVC			FKL			

Figure 4: The performance of different eye movement features on the 382 images for recognizing the pleasantness of the images. In this set, there is no gender agreement on the emotional gist of some of the images and a diverse visual class category exist. Chance level is 33%.

Actual	Prediction			Prediction			Prediction		
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant
	37.1	35.3	27.5	34.8	33.0	32.2	31.4	33.0	35.6
	30.2	32.2	37.6	35.0	32.4	32.6	34.7	32.1	33.2
	31.5	31.8	36.6	30.3	34.5	35.2	33.8	34.7	31.5
mA=35.3%			mA=34.2%			mA=31.7%			
(p = 0.44, 95% CI=32.53–36.12)			(p = 0.59, 95% CI= 32.67-35.65%)			(p = 0.54, 95% CI= 27.85–35.47%)			
all features, 382 images			all features, 296 images			all features, 95 images			
Actual	Prediction			Prediction			Prediction		
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant
	35.5	35.8	28.7	41.8	27.4	30.8	40.2	30.4	29.4
	32.2	30.5	37.3	33.2	33.5	33.2	29.1	32.9	38.0
	29.9	31.2	38.9	28.2	36.8	35.0	30.6	37.2	32.2
mA=34.9%			mA=36.8%			mA=35.1%			
(p = 0.16, 95% CI=33.05–36.86)			(p = 0.09, 95% CI= 33.68–38.65%)			(p = 0.48, 95% CI= 29.45–40.88%)			
FDM + HSS, 382 images			FDM + HSS, 296 images			FDM + HSS, 95 images			

Figure 5: The performance of the fusion of different eye movement-based features. The late fusion of all features and the combination of fixation density map (FDM) and the histogram of saccde slope (HSS) are presented. Chance level is 33%.

Actual	Prediction		
	unpleasant	neutral	pleasant
unpleasant	32.4	31.0	36.6
neutral	33.3	33.3	33.3
pleasant	33.9	34.8	31.3

mA=32.3% ( $p = 0.73$ , 95% CI=28.81–35.85%)  
entropy of fixation density map 95 images

Actual	Prediction		
	unpleasant	neutral	pleasant
unpleasant	33.4	33.1	33.4
neutral	33.8	32.8	33.5
pleasant	32.5	34.5	33.0

mA=33.1% ( $p = 0.86$ , 95% CI=32.17–33.93%)  
entropy of fixation density map 296 images

Actual	Prediction		
	unpleasant	neutral	pleasant
unpleasant	32.3	32.6	34.4
neutral	33.1	33.6	33.3
pleasant	33.9	33.9	32.2

mA=32.9% ( $p = 0.73$ , 95% CI=30.09–34.81%)  
entropy of fixation density map 382 images

Figure 6: Location independent fixation information: the performance of entropy of fixation density map.

To further look into this phenomenon, we conduct an experiment by varying the number of observers and trying to predict the pleasantness for the 296 images. This is inspired by the observer consistency concept [65], used for predicting the number of observers suitable for saliency prediction. The hypothesis is that a minimum number of observers are needed to obtain a meaningful scene representation in order to achieve above chance performance in pleasantness recognition. We summarize the results in terms of the mean average accuracy and confidence intervals in Figure 7. It reveals that at least 15 observers are needed to achieve above chance classification.

### 5.3. Visual Features

We assessed the pleasantness detection from visual features in the three category (unpleasant, neutral, and pleasant) scenario. Similar to the experiments for eye movements, we use the three sets of images consisting of 95, 296, and 382 images. The visual features, which are of 4096 dimension, are L1-normalized and fed to the linear SVM. The results are summarized in Figure 8. In general, the average performance of visual features are in favor of deep features.

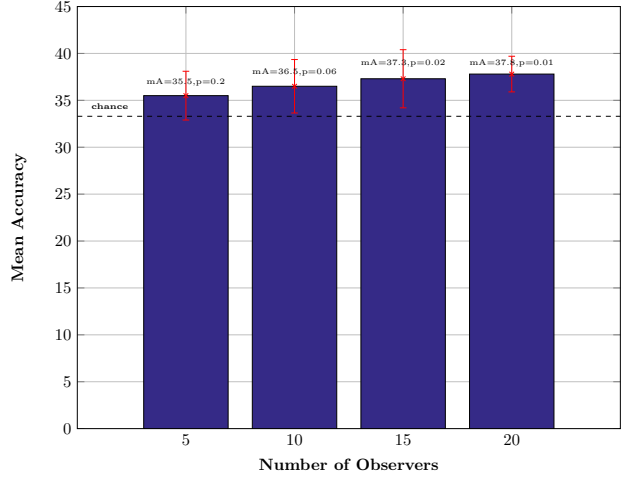


Figure 7: The effect of number of observers on performance of fixation density map using the 296 images setup.

The PHOW features as reported in [6] are not performing better than chance on any of the image sets. The performance for the deep features, however, is better than that of the PHOW features. It is, however, worth noting that, considering the 95 images of [6], which consists of only one visual class category of “People and Daily Activity”, the performance is not better than chance for both features.

By extending the image set to include more images, the deep features perform significantly above chance, that is 43.1% [ $p < 0.001$ , 95% CI=40.59–45.52] and 42.2% [ $p < 0.001$ , 95% CI=38.89–46.85], for 296 images and 382 images using VGG16, respectively. The results for VGG19 are also similar, 43.9% [ $p < 0.001$ , 95% CI=42.17–46.83] and 41.8% [ $p < 0.001$ , 95% CI=38.99–45.67] for 296 and 382 images.

Interestingly, the behaviour of deep features is similar to fixation density maps. We hypothesize that at least one reason behind such a performance behaviour potentially lies on the distribution of emotional categories and visual categories of the images, which is addressed in the following section.

### 5.4. Visual Categories

In this experiment, we explicitly investigate the supporting evidence for the role of visual categories by using visual class category detectors to detect

Actual	Prediction			Prediction			Prediction			95 images
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
	30.9	30.9	38.2	40.7	28.9	30.4	39.5	30.8	29.7	
	31.6	34.7	33.7	29.4	36.5	34.0	32.8	34.3	32.8	
	38.0	34.8	27.2	30.1	34.4	35.4	27.9	34.8	37.3	
mA=30.9%			mA=37.6%			mA=37.0%				
(p = 0.37, 95% CI= 28.03–34.98%)			(p = 0.13, 95% CI=30.92–44.08%)			(p = 0.18, 95% CI=30.62–43.37%)				
PHOW			VGG16			VGG19				
Actual	Prediction			Prediction			Prediction			296 images
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
	36.1	34.6	29.3	49.9	27.8	22.3	47.1	29.1	23.8	
	31.7	33.7	34.6	21.9	36.8	41.3	27.6	43.1	29.3	
	32.5	31.9	35.7	22.7	37.2	40.1	25.0	29.5	45.5	
mA=35.1%			mA=43.1%			mA=43.9%				
(p = 0.26, 95% CI=32.74–37.47%)			(p < 0.001, 95% CI=40.59–45.52%)			(p < 0.001, 95% CI=42.17–46.83%)				
PHOW			VGG16			VGG19				
Actual	Prediction			Prediction			Prediction			382 images
	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	unpleasant	neutral	pleasant	
	34.0	36.1	29.9	48.9	28.0	23.0	47.3	28.6	24.1	
	32.4	32.8	34.8	23.6	37.1	39.3	23.2	37.8	39.0	
	33.6	31.4	35.1	22.8	36.5	40.7	24.6	35.2	40.2	
mA=34.0%			mA=42.2%			mA=41.8%				
(p = 0.63, 95% CI=31.89–36.10%)			(p < 0.001, 95% CI=38.89–46.85%)			(p < 0.001, 95% CI=38.99–45.67%)				
PHOW			VGG16			VGG19				

Figure 8: The performance of visual features for different number of images.

hidden visual categories of images. We apply an ImageNet [66] pre-trained model of object recognition [57] on the images to obtain the score for each of the 1000 classes of ImageNet making a feature vector of 1000 dimensions. Then, we train a classifier using the score of the visual class categories of each image as feature, akin to classemes [67]. The results, summarized in Figure 9, reveal that the visual class categories are strong pleasantness predictors when there exists enough visual class category diversity. Nonetheless, in the case of the 95 images, which consist of one hand-labelled class of “People & Daily Activity”, they fail performing significantly better than chance. This result consists with the performance of the fixation density maps and their role as visual scene descriptors.

To further look into the role of the visual class categories, by assigning the visual class label of the maximum score to each image, we can detect 61 visual classes within 95 images, 202 visual class categories within the 296 images, and 240 visual classes within 382 images. We visualize the detected ImageNet visual category and their ground truth emotional class category in Figure 10. These visualizations clearly indicate that some visual classes are emotionally dif-

ferent from each other and detecting visual categories can facilitate emotion prediction as there seems to be a strong bias towards different emotions in the visual categories. For example, based on this analysis on 382 images, “cock” is absolutely neutral, while a “tarantula” is definitely unpleasant. There are also visual class categories of mixed valence such as “maillot”, which is either neutral or pleasant. This latter example also signifies the role of agreement on pleasantness, as the visual class category of “maillot” does not exist in the 296 images where both visual features and fixation density maps perform best.

### 5.5. Eye Movements and Deep Features

To assess the potential gain obtained by combining the eye movement and deep image features, we study the combination of eye movements and deep visual features in a late fusion scheme. The following combinations are studied: the visual features and fixation density maps, and the visual features and all eye movement features. For the 296 images, the obtained mean accuracy is 43.3% [ $p < 0.001$ , 95% CI=41.3–46.68] and 39.8% [ $p < 0.001$ , 95% CI=37.55–42.66], respectively. The confusion matrices are summarized

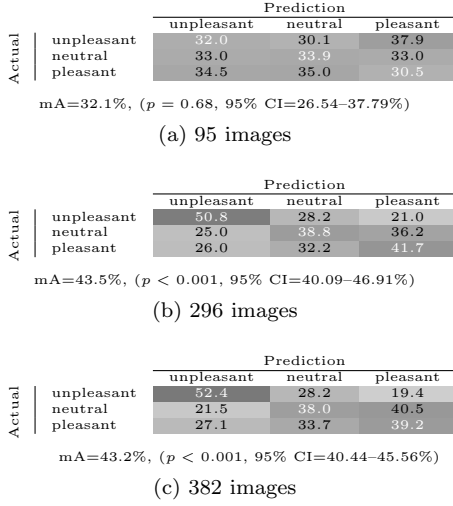


Figure 9: The performance of detection from class categories using VGG16 detections: confusion matrix.

in Figure 11. Similar results hold for the 382 images, where the mean accuracy is 42.5% [ $p < 0.001$ , 95% CI=40.2–46.11] and 39.3% [ $p < 0.001$ , 95% CI=37.51–41.31] for visual features and fixation density maps, and visual features and all eye movement features, respectively. Since the eye movement signals of the current data are not that informative in our framework, it is not surprising that the combination of features also does not improve much over the deep features.

## 6. Discussion

Affective multimedia tagging is a challenging task. It becomes substantially difficult when dealing with images rather than videos. In the case of videos, the audio combined with the visual features make a strong cue for determining the emotional message of a scene. On the other hand, the affective tagging of a still image mostly relies only on visual features and the physiological signals of the observers. In this study, we focused on visual features and eye movements as cues for detecting the pleasantness of still images. The emphasis, however, is on the recording of the eye movements mimicking a setup of everyday user experience using an affordable eye tracking

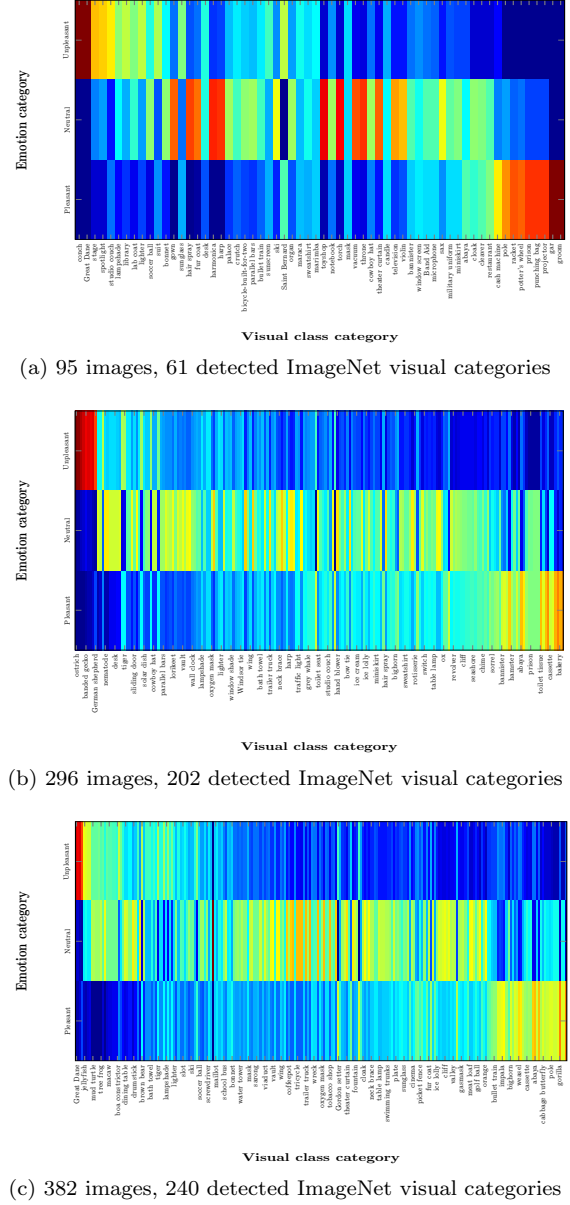


Figure 10: Visual class categories detected with the VGG16 network and their valence according to emotional ratings. It seems there is a strong bias in visual class categories, towards a specific valence value. Some class categories are more pleasant and some are more unpleasant, while some class categories are having multiple ratings. The red color indicates higher probability. From top to bottom, the image emotion changes from unpleasant to pleasant. Selected visual class names are shown for better visualization.

		Prediction		
		unpleasant	neutral	pleasant
Actual	unpleasant	51.3	26.1	22.6
	neutral	20.6	38.8	40.5
	pleasant	23.0	37.1	39.8

mA=43.3% ( $\kappa = 0.37$ , 95% CI=41.3–46.68%)

fixation density map + VGG16

		Prediction		
		unpleasant	neutral	pleasant
Actual	unpleasant	44.2	30.4	25.4
	neutral	27.2	36.1	36.7
	pleasant	26.7	34.1	39.2

mA=39.8% ( $p < 0.001$ , 95% CI=37.55–42.66%)

all eye features + VGG16

Figure 11: The combination of visual features and eye movements for the 296 images: confusion matrices.

device. Consequently, the current data is difficult because it potentially involves a higher level of noise compared with the previous datasets.

The proposed approach and features from eye movements are able to decode image pleasantness only in one setup. The experiment arrangement plays a major role in this respect because we are using a less restrictive setup. Compared to a controlled setup, we are not using a dim room and do not use a chin rest, while the user performs calibration himself/herself after receiving training, consistent with the HCI concept of affordable eye tracking device development. We are using an affordable eye tracking device that has a minimum sampling rate of 60Hz. Not having any control over selecting a fixed sampling rate, makes the precision of fixation durations, and any property relating to time, vulnerable to noise. It is worth noting that we are using the API provided by the manufacturer to obtain eye movement features, e.g., fixations, fixation durations, and gaze points. This can be alerting in the sense that scaling up the utilization of the inference from eye movement using current affordable technologies might be challenging if not impossible. Furthermore, we also tried optimizing the feature parameters, e.g., the number of bins, etc. via cross-validation. Nonetheless, the results were not significantly different for various parameter values.

On the use of computer-vision-based techniques, we focused mostly on techniques that analyze the image content rather than the users’ facial expressions. While the latter potentially provide a better

user-dependent emotion estimation, that can be aggregated to infer a user-independent majority vote tag, we would like to rely on techniques that facilitate easier concealing of the observer identity and provide a means of unobtrusive signal crowd-sourcing. Thus, facial expression analysis was avoided in this study. It is worth noting that although an eye tracker uses a camera, the hardware can be tweaked not to keep track of the faces, which alleviates the privacy concerns.

In this study, we found that visual categories are a strong predictor for user-independent pleasantness recognition of images. This finding is indeed well aligned with the studies promoting that image class categories are emotionally different from each other [68]. It necessitates us to revisit the computer vision techniques, particularly those developed by adapting pre-trained deep CNNs of image classification tasks like [32, 33], and to study the emotional bias of visual class categories in the databases. Unfortunately, such a study goes beyond the current manuscript. We, thus, will address it separately in future work.

The successful application of eye movements as the sole inference medium depends on many factors including the sensor, the experiment setup and the employed machine learning techniques. Using only the data from an affordable eye tracker, we could fairly well decode the image pleasantness in a subset of images. Considering our less constrained setup, which resembles a scenario in which a naive user interacts with a computer freely, we are slightly alerted that the wide-spread utilization of some signals, such as eye movements, may not be as easy as the community may have expected. Despite some improvements, e.g. [69, 48, 49], we still need robust and vigorous machine learning techniques and more reliable sensors. To conclude, it seems that inference from eye movements is difficult using current sensors and techniques. The success of a method can also be severely affected by the level of control in the interaction and experiment, which is alerting for HCI interfaces in daily use. Even in controlled experiments, it is worth noting that the controversy may exist (e.g., check [47] versus [70]).

The data set is available at <https://github.com/>



## 7. Conclusions

In this study, we investigated the performance of eye movements for valence recognition in natural images in a less restrictive setup by introducing a new data set. An affordable eye tracker (more precisely a Tobii EyeX controller) was put into test and several approaches were studied. The results are promising in one specific case, albeit not compelling enough. This alerts us to be more careful while hoping the wide-spread use of eye tracking-based inferences in the HCI paradigm soon, at least for image pleasantness recognition.

We also studied visual features based on the deep CNNs. While the traditional features failed for pleasantness detection using natural images, deep CNNs were outperforming all the features and provided the top performance in our experiments. Evaluating the visual features further, we revealed that the visual class categories are indeed strong valence predictors. Although such a phenomenon can help boosting the algorithms that are based on visual features, it also necessitates revisiting the results of the existing methods in a bias-free setting.

## 8. Acknowledgement

This work was supported by the Finnish Center of Excellence in Computational Inference Research (COIN).

## References

## References

- [1] O. Barral, M. J. Eugster, T. Ruotsalo, M. M. Spapé, I. Kosunen, N. Ravaja, S. Kaski, G. Jacucci, Exploring peripheral physiology as a predictor of perceived relevance in information retrieval, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, ACM, New York, NY, USA, 2015, pp. 389–399. doi:10.1145/2678025.2701389.
- [2] V. Yanulevskaya, J. van Gemert, K. Roth, A. Herbold, N. Sebe, J. Geusebroek, Emotional valence categorization using holistic image features, in: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, 2008, pp. 101–104. doi:10.1109/ICIP.2008.4711701.
- [3] J. Machajdik, A. Hanbury, Affective image classification using features inspired by psychology and art theory, in: Proceedings of the international conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 83–92.
- [4] S.-B. Cho, Emotional image and musical information retrieval with interactive genetic algorithm, Proceedings of the IEEE 92 (4) (2004) 702–711. doi:10.1109/JPROC.2004.825900.
- [5] W. Wang, Q. He, A survey on emotional semantic image retrieval, in: 15th IEEE Int. Conf. on Image Processing, 2008, pp. 117–120.
- [6] H. Tavakoli, V. Yanulevskaya, E. Rahtu, J. Heikkilä, N. Sebe, Emotional valence recognition, analysis of salience and eye movements, in: ICPR, 2014.
- [7] J. Itten, The art of color : the subjective experience and objective rationale of color, John Wiley, New York, 1973.
- [8] C. Colombo, A. Del Bimbo, P. Pala, Semantics in visual information retrieval, Multimedia, IEEE 6 (3) (1999) 38–53. doi:10.1109/93.790610.
- [9] L.-C. Ou, M. R. Luo, A. Woodcock, A. Wright, A study of colour emotion and colour preference. Part I: Colour emotions for single colours, Color Research & Application 29 (3) (2004) 232–240. doi:10.1002/co1.20010.
- [10] P. Lang, M. Bradley, B. Cuthbert, International affective picture system (IAPS): Affective ratings of pictures and instruction manual, Tech. Rep. A-8, University of Florida, Gainesville, FL (2008).

- [11] P. Lang, The emotion probe: Studies of motivation and attention, *American psychologist* 50 (1995) 372–372.
- [12] M. M. Bradley, S. Hamby, A. Lw, P. J. Lang, Brain potentials in perception: Picture complexity and emotional arousal, *Psychophysiology* 44 (3) (2007) 364–373. doi:10.1111/j.1469-8986.2007.00520.x.
- [13] M. M. Bradley, M. Codispoti, D. Sabatinelli, P. Lang, Emotion and motivation II: sex differences in picture processing, *Emotion* 1 (3) (2001) 276–298.
- [14] R. Arnheim, *Art and visual perception: A psychology of the creative eye*, 1974.
- [15] M. Bar, M. Neta, Humans prefer curved visual objects, *Psychological Science* 17 (8) (2006) 645–648.
- [16] L. Nummenmaa, J. Hyönä, M. G. Calvo, Eye movement assessment of selective attentional capture by emotional pictures, *Emotion* 6 (2) (2006) 257–268.
- [17] K. Humphrey, G. Underwood, T. Lambert, Salience of the lambs: A test of the saliency map hypothesis with pictures of emotive objects, *Journal of Vision* 12 (1). arXiv:<http://www.journalofvision.org/content/12/1/22.full.pdf+html>, doi:10.1167/12.1.22.
- [18] Y. Niu, R. M. Todd, M. Kyan, A. K. Anderson, Visual and emotional salience influence eye movements, *ACM Trans. Appl. Percept.* 9 (3) (2012) 13:1–13:18. doi:10.1145/2325722.2325726.
- [19] W.-N. Wang, Y.-L. Yu, Image emotional semantic query based on color semantic description, in: *Machine Learning and Cybernetics*, 2005. Proceedings of 2005 International Conference on, Vol. 7, 2005, pp. 4571–4576 Vol. 7.
- [20] M. Solli, R. Lenz, Color emotions for image classification and retrieval, in: *CGIV*, 2008, p. 367371.
- [21] M. Solli, R. Lenz, Color based bags-of-emotions, in: *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns, CAIP '09*, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 573–580.
- [22] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, N. Sebe, In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings, in: *Proceedings of the 20th ACM international conference on Multimedia, MM '12*, ACM, New York, NY, USA, 2012, pp. 349–358.
- [23] A. Sartori, D. Culibrk, Y. Yan, N. Sebe, Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings, in: *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 311–320.
- [24] A. Bovik, M. Clark, W. Geisler, Multichannel texture analysis using localized spatial filters, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12 (1) (1990) 55–73. doi:10.1109/34.41384.
- [25] A. Vailaya, M. A. Figueiredo, A. K. Jain, H.-J. Zhang, Image classification for content-based indexing, *Trans. Img. Proc.* 10 (1) (2001) 117–130. doi:10.1109/83.892448.
- [26] X. Lu, P. Suryanarayan, R. B. Adams, Jr., J. Li, M. G. Newman, J. Z. Wang, On shape and the computability of emotions, in: *Proceedings of the 20th ACM international conference on Multimedia, MM '12*, ACM, New York, NY, USA, 2012, pp. 229–238. doi:10.1145/2393347.2393384.
- [27] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, *ACM Trans. Multimedia Comput. Commun. Appl.* 2 (1) (2006) 1–19. doi:10.1145/1126004.1126005.
- [28] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neu-

- ral networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., 2012, pp. 1097–1105.
- [29] M. Koskela, J. Laaksonen, Convolutional network features for scene recognition, in: *Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14*, ACM, New York, NY, USA, 2014, pp. 1169–1172. doi:10.1145/2647868.2655024.
- [30] C. Szegedy, A. Toshev, D. Erhan, Deep neural networks for object detection, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, 2013, pp. 2553–2561.
- [31] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *CVPR (to appear)*arXiv:1411.4038.
- [32] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, AAAI Press, 2015, pp. 381–388.
- [33] J. Wang, J. Fu, Y. Xu, T. Mei, Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks, in: *IJCAI*, 2016.
- [34] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [35] R. Subramanian, V. Yanulevskaya, N. Sebe, Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements, in: *Proceedings of the 19th ACM international conference on Multimedia, MM '11*, ACM, New York, NY, USA, 2011, pp. 33–42. doi:10.1145/2072298.2072305.
- [36] R. Subramanian, H. Katti, R. Huang, T.-S. Chua, M. Kankanhalli, Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis, in: *Proceedings of the 17th ACM international conference on Multimedia, MM '09*, ACM, New York, NY, USA, 2009, pp. 729–732. doi:10.1145/1631272.1631399.
- [37] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, *Affective Computing, IEEE Transactions on* 3 (2) (2012) 211–223. doi:10.1109/T-AFFC.2011.37.
- [38] D. Lowe, Object recognition from local scale-invariant features, in: *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Vol. 2, 1999, pp. 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410.
- [39] H. R.-Tavakoli, A. Atyabi, A. Rantanen, S. J. Laukka, S. Nefti-Meziani, J. Heikkilä, Predicting the valence of a scene from observers eye movements, *PLoS ONE* 10 (9) (2015) e0138198. doi:10.1371/journal.pone.0138198.
- [40] M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1) (1994) 49 – 59.
- [41] H. A. Wadlinger, D. M. Isaacowitz, Positive mood broadens visual attention to positive stimuli, *Motivation and Emotion* 30.
- [42] J. Tichon, T. M. G. Wallis, T. Visser, S. Riek, Using pupillometry and electromyography to track positive and negative affect during flight simulation, *Aviation Psychology and Applied Human Factors* 4 (1).
- [43] J. Simola, K. L. Fevre, J. Torniaainen, T. Baccino, Affective processing in natural scene viewing: Valence and arousal interactions in eye-fixation-related potentials, *NeuroImage* 106 (2015) 21 – 33. doi:http://dx.doi.org/10.1016/j.neuroimage.2014.11.030.
- [44] C. Lithari, C. Frantzidis, C. Papadelis, A. Vivas, M. Klados, C. Kourtidou-Papadeli, C. Pappas,

- A. Ioannides, P. Bamidis, Are females more responsive to emotional stimuli? a neurophysiological study across arousal and valence dimensions, *Brain Topography* 23 (1) (2010) 27–40. doi:10.1007/s10548-009-0130-5.
- [45] A. Torralba, A. Oliva, M. Castelhano, J. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, *Psychol Rev* 113 (4) (2006) 766–86.
- [46] B. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions., *Journal of Vision* 14 (7).
- [47] M. R. Greene, T. Liu, J. M. Wolfe, Reconsidering Yarbus: A failure to predict observers task from eye movement patterns, *Vision Research* 62 (2012) 1 – 8. doi:10.1016/j.visres.2012.03.019.
- [48] A. Borji, L. Itti, Defending yarbus: Eye movements reveal observers task, *Journal of Vision* 14 (5).
- [49] C. Kanan, N. A. Ray, D. N. F. Bseiso, J. H. Hsiao, G. W. Cottrell, Predicting an observer’s task using multi-fixation pattern analysis, in: *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA ’14*, ACM, New York, NY, USA, 2014, pp. 287–290. doi:10.1145/2578153.2578208.
- [50] L. van der Maaten, Learning discriminative fisher kernels, in: *International Conference on Machine Learning (ICML)*, 2011, pp. 217–224.
- [51] O. Le Meur, T. Baccino, A. Roumy, Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking, in: *ACM Multimedia, Phoneix, United States*, 2011.
- [52] M. Mancas, O. L. Meur, Memorability of natural scenes: The role of attention, in: *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*, 2013, pp. 196–200.
- [53] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision* 60 (2) (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.
- [54] A. Bosch, A. Zisserman, X. Muoz, Image classification using random forests and ferns, in: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8. doi:10.1109/ICCV.2007.4409066.
- [55] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, Speeded-up robust features (SURF), *Computer Vision and Image Understanding* 110 (3) (2008) 346 – 359, similarity Matching in Computer Vision and Multimedia. doi:http://dx.doi.org/10.1016/j.cviu.2007.09.014.
- [56] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating color descriptors for object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (9) (2010) 1582–1596.
- [57] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *ICLR*, 2015.
- [58] C. Cortes, V. Vapnik, Support-vector networks, *Machine Learning* 20 (3) (1995) 273–297. doi:10.1023/A:1022627411411.
- [59] H. He, E. A. Garcia, Learning from imbalanced data, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 21 (9) (2009) 1263 – 1284.
- [60] A. Atyabi, M. Luerssen, S. Fitzgibbon, D. M. W. Powers, Evolutionary feature selection and electrode reduction for eeg classification, in: *Evolutionary Computation (CEC), 2012 IEEE Congress on*, 2012, pp. 1–8. doi:10.1109/CEC.2012.6256130.
- [61] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* 12 (2) (1947) 153 – 157.

- [62] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* 42 (2001) 145–175.
- [63] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2) (2007) 300–312.
- [64] P. Gomez, A. von Gunten, B. Danuser, Content-specific gender differences in emotion ratings from early to late adulthood, *Scandinavian Journal of Psychology* doi:10.1111/sjop.12075.
- [65] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, Tech. Rep. MIT-CSAIL-TR-2012-001, Massachusetts institute of technology (2012).
- [66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: *CVPR09*, 2009.
- [67] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classemes, in: *Proceedings of the 11th European Conference on Computer Vision: Part I, ECCV’10*, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 776–789.
- [68] M. Bradley, P. J. Lang, The international affective picture system (IAPS) in the study of emotion and attention, in: J. A. Coan, J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment*, Oxford University Press, 2007, pp. 29–46.
- [69] H. Zhang, M. Gnen, Z. Yang, E. Oja, Understanding emotional impact of images using bayesian multiple kernel learning, *Neurocomputing* 165 (2015) 3–13. doi:http://dx.doi.org/10.1016/j.neucom.2014.10.093.
- [70] A. Haji-Abolhassani, J. J. Clark, An inverse yabus process: Predicting observers task from eye movement patterns, *Vision Research* 103 (2014) 127–142. doi:http://dx.doi.org/10.1016/j.visres.2014.08.014.